The promise and the peril of using social influence to reverse harmful traditions

Efferson, Charles ; Vogt, Sonja ; Fehr, Ernst

Abstract: For a policy-maker promoting the end of a harmful tradition, conformist social influence is a compelling mechanism. If an intervention convinces enough people to abandon the tradition, this can spill over and induce others to follow. A key objective is thus to activate such spillovers and amplify an intervention's effects. With female genital cutting as a motivating example, we develop empirically informed analytical and simulation models to examine this idea. Even if conformity pervades decision-making, spillovers can range from irrelevant to indispensable. Our analysis highlights three considerations. First, ordinary forms of individual heterogeneity can severely limit spillovers, and understanding the heterogeneity in a population is essential. Second, although interventions often target samples of the population biased towards ending the harmful tradition, targeting a representative sample is a more robust way to achieve spillovers. Finally, if the harmful tradition contributes to group identity, the success of spillovers can depend critically on disrupting the link between identity and tradition.

DOI: https://doi.org/10.1038/s41562-019-0768-2

The promise and the peril of using social influence to reverse harmful traditions

Charles Efferson[1,*], Sonja Vogt[2,3,5], and Ernst Fehr[4,6]

[1] HEC Lausanne, University of Lausanne, Switzerland

[2] Department of Social Sciences, University of Bern, Switzerland

[3] Centre for Development and Environment, University of Bern, Switzerland

[4] Department of Economics, University of Zurich, Switzerland

[5] Centre for Experimental Social Sciences, Nuffield College, University of Oxford, U.K.

[6] Center for Child Well-Being and Development, University of Zurich, Switzerland

[*] Address correspondence to CE (charles.efferson@unil.ch)

For a policy maker promoting the end of a harmful tradition, conformist social influence is a compelling mechanism. If an intervention convinces enough people to abandon the tradition, this can spill over and induce others to follow. A key objective is thus to activate spillovers and amplify an intervention's effects. With female genital cutting as a motivating example, we develop empirically informed analytical and simulation models to examine this idea. Even if conformity pervades decision making, spillovers can range from irrelevant to indispensable. Our analysis highlights three considerations. First, ordinary forms of individual heterogeneity can severely limit spillovers, and understanding the heterogeneity in a population is essential. Second, although interventions often target biased samples of the population, targeting a representative sample is a more robust approach to spillovers. Finally, if the harmful tradition contributes to group identity, spillovers can hinge critically on disrupting the link between identity and tradition.

Introduction

Harmful traditions create a basic conflict[1]. They reflect the values and traditions of their respective cultures[2,3], and tolerance of cultural differences implies some degree of acceptance. In contrast, a commitment to cross-cultural standards like universal human rights[4] and the idea that cultures can evolve in destructive ways[5] suggests exactly the opposite response. A policy maker who favours the latter view can intervene to disrupt a harmful tradition and steer cultural change in an alternative direction. In doing so, however, the policy maker stands in direct opposition to the norms and values of the target population[3]. This is the basic conflict of applied cultural evolution.

Conformist social influence provides an appealing mechanism for managing this conflict, and it has generated considerable interest as a policy tool in various domains[1,6–8]. Specifically, if decision makers have an interest in behaving like those around them, a policy maker can potentially recruit this individual-level mechanism to amplify the effects of an intervention at the aggregate level. In such cases, the intervention has a direct effect that leads some agents to change because of exposure to the intervention proper. An indirect effect also obtains if the initial change among a subset of agents provokes others to follow without further interference. We refer to these indirect effects as "spillovers".

To the extent that spillovers are responsible for changes in behaviour, change is endogenous. Endogenous change limits the need for the policy maker to meddle directly in the culture of the target population. This could reduce the cost of reversing a harmful tradition, increase the perceived legitimacy of change, reduce the tendency for people to conclude their culture is under attack, and reduce inter-cultural strife

and the associated risk of backlash. Spillovers hold much appeal for these reasons, and their potential informs policy related to female genital cutting[2,3,9–12], child marriage[3,13–15], open defecation[11,16], domestic violence[12,17], and a preference for sons[12]. Research has also highlighted the role of social influence, and in some cases its policy relevance, with respect to smoking[18], foot binding[19], alcohol consumption[20], duelling[21], obesity[22], bullying behaviour[23], energy conservation[24], tax compliance[25], and freshwater conservation[26,27]. In the following, with female genital cutting as a motivating example, we examine both the potential and the potential limitations of spillovers. In particular, we detail why aggregate-level spillovers, a prominent objective of programmes promoting the abandonment of cutting[3,12], may or may not follow from conformist social influence at the individual level.

Cutting takes various forms, but it is often a serious and potentially dangerous procedure that involves significant tissue removal and possibly infibulation[28]. The scale of the practice is also daunting. Current estimates place the number of cut females in the world at far beyond 100 million, with an estimated three million girls at risk of cutting each year[28]. Because the scale is so vast, endogenous spillovers may in fact be necessary to end female genital cutting, and cutting provides an archetypical example of how policy makers hope to use spillovers to shape cultural evolution in ways consistent with their objectives[3,10].

Importantly, however, recent empirical studies on cutting suggest a fundamental caveat. Many families seem to have an interest in conforming to the local practice in terms of whether or not they cut their daughters[3,29–33]. This kind of conformist social influence is an important mechanism for generating spillovers, which suggests that programmes promoting the abandonment of cutting are right to consider potential spillovers. Even so, attitudes and practices related to cutting are quite heterogeneous at extremely local scales[12,29–31,34]. As we show below, this mix implies that conformist decision making at the individual level, though present, may not be creating opportunities to trigger significant spillovers[12,30,31]. With this as our point of departure, we develop several models to examine the following key question. How do various forms of heterogeneity combine with an overall emphasis on conformity to affect the scope for spillovers?

Before turning to these models, we present a simple illustration to cultivate intuition. Assume a population of agents. Think of these agents as decision-making parents who decide whether or not to cut. We treat the family as a single decision-making unit with a unified position at a given point in time. A family may change its position through time, but at any given point in time the family is either a cutting family or a non-cutting family. Parents in a cutting family cut any daughters, and they socialise any sons to value cut wives. Parents in a non-cutting family do not cut their daughters, and they socialise any sons to value uncut wives. In this way, we categorize a family as cutting or non-cutting regardless of the exact composition of girls and boys. This assumption is consistent, in particular, with recent research showing that parents have the same view of cutting in reference to both their daughters and the wives of their sons[35].

Each family is subject to social influence in the sense that the family is increasingly likely to abandon cutting as the proportion of families who do not cut increases. Aside from this common interest in behaving like others, families vary. Consistent with experimental studies on conformity[36–39] and with data on cutting[3,31,40], families differ in terms of how strongly they respond to information about the behaviour of others (Fig. 1a). We assume two types of family, a type that responds strongly to information about others and a somewhat less responsive type. As a thought exercise, imagine we can manipulate the proportion of families of the less responsive type. As the less responsive type becomes more common, the stable equilibria converge on a uniform mix of cutting and non-cutting families. This means cutting practices become increasingly heterogeneous and less norm-like in equilibrium (Fig. 1b). By extension, the potential to incite the endogenous abandonment of cutting via spillovers diminishes and eventually disappears altogether (Fig. 1b). As we show below with more realistic models, this kind of generic result arises often. Social influence may be pervasive, but the scope for inciting beneficial spillovers can vary tremendously when people are heterogeneous in other ways.

We begin with a simple model that assumes everyone is the same, and we progressively incorporate different forms of heterogeneity known to be important. A longstanding hypothesis is that families face incentives to coordinate their choices related to cutting. These incentives create a situation in which people conform, and spillovers are thus possible[19]. Coordination incentives related to cutting can take various forms[9], but empirical studies indicate that marriageability concerns are important[3,31,33,41]. To illustrate, consider a society in which people cut their daughters to signal that these girls will one day become morally upright and sexually faithful wives[41]. Any given family in such a society faces powerful incentives to cut its daughters and to value cut wives for its sons. Because of the shared understanding of what cutting signals, a family's best option is to cut so that other families perceive the family's daughters as good potential wives and mothers. The family's incentives also favour cut wives for its sons. Otherwise, the family would invite the perception that its sons are destined to raise the children of other men.

In contrast, imagine a society in which people do not associate cutting with sexual fidelity. Families have no incentive to cut their daughters, and they have no incentive to demand uncut wives for their sons. In effect, families can avoid the health risks of cutting without signalling that their daughters are untrustworthy. Moreover, uncut wives for a family's sons ensure that daughters-in-law give birth without undue complications, granddaughters are themselves likely to avoid risks associated with cutting, and all of this without the family's sons carrying the stigma of presumed cuckolds.

Coordination incentives mean that, when parents consider the future husbands and wives of their daughters and sons, the expected values of cutting and not cutting vary according to how common the practices are[19,32]. If cutting is sufficiently common, cutting is a family's best choice because the probability of a future

marriage with a cutting family is relatively high. If not cutting is sufficiently common, the opposite holds. The transition between these two cases occurs at an indifference point. An indifference point is the specific proportion of non-cutting families that renders a focal family indifferent between cutting and not cutting. At this indifference point, the expected value of cutting is the same as that of not cutting.

When preferences related to the intrinsic value of cutting are homogeneous, every family has the same indifference point, which we label $\tilde{q}$. To generate cultural change, assume each family evaluates its current practice occasionally and updates if necessary. Let q_t be the proportion of families not cutting at time t. A family who updates chooses with certainty the option that maximises expected payoffs given the current distribution of behaviours in the population. If the proportion choosing not to cut is less than $\tilde{q}$ but greater than zero, updating families choose to cut, and the proportion not cutting declines through time (i.e. for some $t' > t$, $q_{t'} < q_t$). If the proportion not cutting is greater than $\tilde{q}$ but less than one, updating families choose not to cut, and the proportion not cutting rises through time (i.e. for some $t' > t$, $q_{t'} > q_t$).

This model holds clear implications for a policy maker who wants to promote the abandonment of cutting. If not cutting is sufficiently rare, then cutting has the highest expected value, and families choose accordingly. The population converges to a self-reinforcing equilibrium in which everyone cuts. Coordination incentives work against the policy objective. However, if not cutting is sufficiently common, then not cutting has the highest expected value. The population converges to a self-reinforcing equilibrium in which no one cuts. Coordination incentives work in favour of the policy objective. The common indifference point, $\tilde{q}$, separates these two regimes. $\tilde{q}$ is itself an equilibrium, but it is not stable. Any small deviation from $\tilde{q}$ sets the population on a path towards either widespread cutting or widespread abandonment, and for this reason $\tilde{q}$ is sometimes called a "tipping point".

If the policy maker designs and implements an intervention, whatever it may be, that convinces any proportion of families greater than $\tilde{q}$ to abandon cutting, coordination incentives take over and lead all remaining families to stop cutting. This is perhaps the canonical model of how a policy maker can recruit endogenous social forces to do her bidding. Designing an effective intervention can be difficult, but any adequate intervention triggers spillovers that eventually result in a complete transition to abandonment. Once the policy maker has triggered these spillovers, she can move on to some other task even if many families still cut. She has crossed the tipping point, and endogenous forces have taken over.

What if, however, in addition to every family's shared interest in coordinating, families also vary in terms of the intrinsic values[42–44] they attach to cutting? For example, some families may be preoccupied with the health risks of cutting[33]. These families want to coordinate, but they would prefer to coordinate on not cutting as opposed to cutting. Other families may have fully internalised the notion that cutting is necessary for girls to grow up and become morally upright women[33]. These families also want to coordinate, but they

would prefer to coordinate on cutting. All in all, any given family has a ranking over the pure-strategy equilibria of the underlying coordination game and a ranking over the costs of deviating from the different equilibria. Families disagree, however, about these rankings. Recent empirical work on cutting[3,30,31,33,45] has provided strong support for the existence of such preference heterogeneity.

This kind of heterogeneity means that different families, whatever their reasons, have different indifference points. Heterogeneity of this sort can have stark consequences. If F is the cumulative distribution function for indifference points, and if each family evaluates its practice in every period and updates if necessary, the proportion not cutting evolves according to $q_{t+1} = F(q_t)$. This is a classic result[42,46] interpreted here in terms of female genital cutting[12].

Importantly, depending on the shape of F, a tipping point may or may not exist, and this is true even though everyone responds strongly to coordination incentives (Supplementary Information, e.g. Supplementary Figs. 3-14). In particular, by assumption each agent either deterministically follows a sufficiently large majority that cuts or a sufficiently large majority that does not cut. In this sense, all agents are strong conformists. Agents vary, however, in terms of the values of q_t that induce them to switch from one behaviour to the other. We refer to these values as "thresholds", which can be interpreted as indifference points under coordination incentives. The structure of threshold heterogeneity controls whether or not coordination incentives and the associated tendency to conform create a tipping point. This is important because the policy maker's concern does not centre around the existence of coordination incentives per se, but rather around the potential for endogenous cultural change. The former concerns decision making at the individual level; the latter concerns dynamics at the population level.

Heterogeneity shapes the potential for social influence to drive spillovers, and this point raises a number of fundamental policy questions. Accordingly, we add a number of mechanisms known to be important to the basic model (Supplementary Information) and examine how these mechanisms interact with policy choices to affect spillovers. First, we assume that individuals who respond to the intervention change their preferences[12,33] and abandon cutting unconditionally. Their threshold values go to zero, which effectively creates a subset of the population unequivocally committed to abandonment. Recent research shows that an unequivocally committed minority of this sort can tip a population to a new state in which everyone adopts a new opinion or behaviour[47,48]. These are fascinating results based on a game involving pure coordination incentives[49] in which coordinating on any given option is equivalent to coordinating on any other option. In contrast, as explained above, we assume that equilibria can be ranked, and decision makers vary in terms of their rankings. This is consistent with empirical research on cutting[3,12,29–31], and more broadly we suspect that, when harmful traditions involve coordination incentives, associated equilibria will often be ranked. Otherwise, behaviour change would be relatively straightforward precisely because decision makers would

find the policy maker's target equilibrium just as good as the past tradition.

In our case, the intervention creates a committed subpopulation embedded in a larger population of agents with heterogeneous preferences. The intervention thus shifts the threshold distribution in ways that may or may not affect the potential for spillovers, with the details depending on which segment of the population an intervention targets. We consider interventions that target a sample of the population that is relatively amenable to the abandonment of cutting, a randomly selected sample, or a sample relatively resistant to abandonment.

Second, we introduce the assumption that, as agents become increasingly resistant to abandonment, they are decreasingly likely to respond to the intervention. Our own fieldwork in Sudan provides strong evidence for this assumption[33], as does fieldwork from other countries[3]. Third, we introduce various degrees of homophily. Homophily is ubiquitous in human societies[50], and in our case homophily means that agents are most likely to associate with others having initially similar attitudes towards cutting. Finally, we turn to a model in which families link their cultural identities to their cutting practices. This model reflects recent research showing that female genital cutting is, to some extent, a matter of group identity[3,9,12,32]. In all cases, to quantify and standardise endogenous changes in behaviour, we normalise spillovers by scaling the endogenous change that actually occurs by the maximum endogenous change that could occur (Methods).

Results

Spillovers under alternative intervention targets and heterogeneous responses to the intervention

We begin with the case in which everyone is connected, and everyone targeted by the intervention responds in the sense that they abandon cutting unconditionally. With intervention in hand, the policy maker has two decisions to make, the size of the intervention and whom to target (Supplementary Information). Specifically, the policy maker targets a proportion ϕ of the population, and targeted individuals are directly exposed to the intervention. Given ϕ, we assume the policy maker can target amenable agents, randomly selected agents, or resistant agents. Agents amenable to abandonment are amenable in the sense that their pre-intervention thresholds are relatively low. A random sample, in contrast, means the intervention target is not subject to selection bias, and in particular the policy maker selects the target without regard for threshold values. A resistant target, finally, consists of agents with relatively high threshold values.

In terms of generating spillovers, we show analytically (Supplementary Information) that targeting randomly selected agents cannot be worse than targeting amenable agents, and it will often be strictly better. Analogously, targeting resistant agents cannot be worse than targeting randomly selected agents, and it will

often be strictly better. Intuitively, if the policy maker targets amenable agents, she uses her exogenous intervention to accomplish the easiest possible task, which is to convince those agents most amenable to change to abandon cutting. This, in turn, leaves the most difficult possible task to endogenous spillovers, and spillovers are limited as a result. A random sample moderates this problem because it tends to exclude some amenable agents from the intervention and to include some resistant agents. Change via the exogenous intervention might increase in difficulty because of the latter effect, but we ignore this for the moment by assuming that everyone responds to the intervention. This leaves only the effect that a random sample improves conditions for spillovers because it tends to exclude some agents amenable to change. These amenable agents are thus available to abandon cutting via endogenous spillovers. Targeting the most resistant agents continues in the same vein. A resistant target takes the most challenging conceivable task for the intervention, and it leaves the easiest conceivable task for spillovers. Fig. 2 illustrates the logic by showing an example of how the three types of target change F in different ways, with potentially profound consequences for cultural evolution and spillovers.

Fig. 3a presents simulation results that summarise these patterns under a wide array of conditions. In particular, a random target promotes spillovers under a wider range of conditions than an amenable target, and a resistant target promotes spillovers under a wider range of conditions than a random target. These differences are most pronounced when the exogenous intervention is large (e.g. $\phi = 0.5$, Fig. 3a).

Strikingly, however, these differences in spillovers pale in comparison to the differences that arise from variation in pre-existing preferences. Regardless of whether the intervention is small or large, and regardless of the intervention target, the one factor with an overriding effect on spillovers is the distribution of thresholds before the intervention. If this distribution is right skewed (Fig. 3a, above 45° line), most people are amenable to the abandonment of cutting before the policy maker enters the scene, and endogenous spillovers are always pronounced. If the distribution is left skewed (Fig. 3a, below 45° line), most people are initially resistant to abandonment when the policy maker appears, and endogenous spillovers are largely absent. Spillovers are only sensitive to the choices available to the policy maker, namely the size and target of the intervention, when the distribution of threshold values is sufficiently close to symmetric. Other distributions mean the policy maker can promote abandonment via the size of the exogenous intervention, but her choices have surprisingly little effect on endogenous spillovers.

We next add the assumption that agents initially resistant to abandonment are relatively unlikely to respond to the intervention (Supplementary Information). This means that in expectation, with an intervention of size ϕ, some proportion of agents less than ϕ changes due to the intervention. Importantly, a recent field experiment in 122 communities in Sudan found exactly this pattern[33]. The participants most in favour of cutting before the intervention were least likely to respond to the intervention. Adding this mechanism to

a population of heterogeneous agents reduces the scope for a policy maker to influence spillovers (Fig. 2g-2i, Fig. 3b).

In particular, we argued above that, when the pre-existing distribution of preferences neither strongly favours nor strongly disfavours change, the policy maker can improve spillovers by targeting individuals resistant to abandonment. Such a strategy effectively reserves the difficult cases for the intervention and the easy cases for spillovers. This is precisely why targeting resistant agents can foster spillovers. However, if resistant agents do not respond to the intervention, perhaps because budgetary constraints ensure that the intervention is just not enough to convince them, then many of the gains from targeting resistant agents are lost. Indeed, when responses to the intervention are heterogeneous, our results indicate that the policy maker's ability to influence spillovers by choosing the size and target of the intervention may be almost entirely absent (Fig. 3b). This is especially true for moderate and large interventions ($0.2 \leq \phi \leq 0.5$), where spillovers are largely independent of the policy maker's choices. Importantly, this result does not imply that the policy maker cannot influence the number of people who abandon cutting. A large expensive intervention can still convince more families to abandon than a small underfunded intervention via the direct effect of the intervention. The result does imply, however, that the ability to influence spillovers can be surprisingly beyond the policy maker's grasp. This is quite different from results involving a pure coordination game in a population with homogeneous preferences, a setting in which the size of the committed minority is a key consideration[48].

Spillovers in homophilous networks

Adding homophilous social structure erodes the potential for spillovers, but the effects are not monotonic. Adding homophily also reveals the potentially robust advantages of targeting a random sample of the population. Specifically, even moderate degrees of homophily can destroy spillovers under otherwise favourable conditions. For example, consider the case in which everyone targeted by the intervention abandons cutting unconditionally. If everyone is connected and the pre-intervention distribution of thresholds is approximately symmetric, a resistant target often leads to large spillovers (Fig. 3a). Moderate homophily reduces these spillovers substantially (Fig. 4a), and a stronger degree of homophily destroys them entirely (Fig. 4b). Broadly speaking, the damaging effects of homophily are especially noticeable if the intervention targets either an amenable or resistant segment of the population, and a random target often generates spillovers more robustly than either type of biased sample. In effect, because homophily fragments the population to some extent, an intervention directed at a random sample of agents is likely to seed each of the various fragments with some families who abandon cutting. This helps attenuate some of the detrimental effects of homophily on spillovers.

Importantly, however, homophily is not uniformly detrimental. To show precisely how variation in the degree of homophily can affect spillovers, we simulated many different degrees of homophily under a symmetric distribution of threshold values (Beta$(3.375, 3.375)$), which is a case in which the initial distribution of preferences should allow other mechanisms to have clear effects on spillovers. We begin with the case in which all agents targeted by the intervention respond to the intervention. The effects of homophily are complex (Fig. 5). Under a resistant target, increasing homophily decreases spillovers regardless of the size of the intervention (Fig. 5e-5f). In contrast, for small interventions (e.g. $\phi = 0.1$) that target amenable and random samples (Fig. 5a-5d), any degree of homophily is better for spillovers than no homophily at all. That said, even in these cases spillovers tend to reach their maximum value when homophily is weak (e.g. Fig. 5d); further increases in homophily reduce spillovers.

Homophily, in short, has countervailing effects[51]. First, it fragments the population, and so the abandonment of cutting does not necessarily need to be common in the entire population before spillovers take over. Instead, abandonment might only need to gain a foothold in a local fragment, at which point abandonment can spread within the fragment before spreading to other fragments[51,52]. Homophily, however, has a second effect. Fragmentation can also hinder the diffusion of abandonment from one fragment to another[51,53]. For small interventions that target amenable or random samples, the foothold effect dominates (Fig. 5a-5d, $\phi \in \{0.1, 0.2\}$). This finding is consistent with analytical results that effectively assume vanishingly small interventions[51]. Otherwise, limited diffusion between fragments dominates, and increasing homophily leads to declining spillovers (Fig. 5). Overall the analysis also clearly shows that, when homophily is present, random targets generate considerably larger spillovers than amenable or resistant targets (Fig. 5).

When we add the assumption that agents resistant to change are relatively unlikely to respond to the intervention, the effects of homophily are similar. In particular, when comparing some degree of homophily to no homophily, spillovers can sometimes be larger when homophily is present (Fig. 6a-6d). More broadly, however, increasing homophily leads to decreasing spillovers, and in this sense homophily can hinder the abandonment of cutting via endogenous social mechanisms. Indeed, when homophily and heterogeneous responses to the intervention combine, spillovers are typically small or absent altogether (Fig. 6). Whatever the potential for spillovers, the policy maker again does best by targeting a random sample.

Spillovers when harmful traditions are tied to group identity

In 1956 in the Meru district of Kenya, a council of local male leaders, widely seen as the compliant pawns of colonising forces, banned female genital cutting. Over the next three years, thousands of individuals defied the ban, often at great expense to their families, and in many cases girls took matters into their own hands by using razor blades to cut each other[54]. Although the mechanisms at work were no doubt varied and

complex, defiance of the ban was in part an assertion of cultural identity in the face of a threat from an outside influence[54]. The Meru story is perhaps uniquely arresting, but the significance of identity concerns is not. Empirical research indicates that cutting often serves as a means of defining, asserting, and maintaining group identity[2,3,9,12,32,55], and more broadly agents sometimes reject international norms related to human rights to signal their identification with a local culture[56].

We now incorporate identity concerns of this sort by dividing the population into two groups (Methods and Supplementary Information). Each agent responds to the distribution of behaviours in her own group with probability $\beta \in [0.5, 1]$. Conditional on a response to the ingroup, agents tend to conform to the majority choice, and this ingroup conformity can be relatively weak or strong (γ_{ij}, see Methods). With probability $1 - \beta$, an agent responds to the distribution of behaviours in the other group. Conditional on a response to the outgroup, agents tend to adopt the minority choice, and this outgroup anti-conformity can also be relatively weak or strong (μ_{ij}, see Methods). β thus summarises the emphasis agents place on responding to their own group, typically in a conformist fashion, versus responding to the outgroup, typically in an anti-conformist fashion that serves to establish and maintain a distinct group identity.

Analytical results show that, after an intervention of size ϕ, two types of stable equilibria can exist. One type involves the two groups stabilising on exactly the same mix of cutting and non-cutting families. This shared distribution of behaviours represents a balance between responding to the ingroup (β) and responding to the outgroup ($1-\beta$), and the two groups are paradoxically identical in equilibrium. For many equilibria of this sort, the stable proportion of families who do not cut is larger than the exogenous intervention (ϕ), and in this sense spillovers can occur provided the intervention tips the population into the basin of attraction for this equilibrium. That said, an increasing tendency to respond to the outgroup ($\beta \to 0.5^+$) lowers the shared equilibrium proportion of families not cutting, and in this way group identity concerns place important limits on spillovers (Supplementary Information, Supplementary Figs. 27-33, 39, 45-47).

The other type of stable equilibrium involves the two groups converging on two distinct traditions, with one group cutting at a high rate and the other group cutting at a low rate. Equilibria of this sort often appear, all else equal, as responses to the ingroup (β) decrease in frequency, and responses to the outgroup ($1 - \beta$) increase in frequency (Supplementary Information, Supplementary Figs. 33-34, 37-40, 42, 44-46). Equilibria of this sort are fundamentally incompatible with large spillovers because one group commits to the harmful tradition to distinguish itself.

Simulation results with heterogeneous agents in a finite population confirm the general conclusion that an emphasis on group identity is generally devastating for spillovers. Under a restricted set of circumstances, a tendency to respond to the outgroup can actually facilitate spillovers. This effect, however, only occurs when the intervention is small and, conditional on an agent responding to the ingroup, ingroup conformity

is strong (Fig. 7e-7h, $\phi = 0.1$). The overall pattern, however, is one in which an increasing emphasis on responding to the outgroup (declining β) dramatically reduces the potential for a policy maker to precipitate spillovers (Fig. 7). Indeed, as agents approach an equal emphasis on responding to the ingroup versus the outgroup ($\beta = 0.5$), substantial proportions of agents who could abandon cutting endogenously do not do so. Thus, a policy maker confronting a strong link between a harmful tradition and group identity should develop strategies for destabilizing this link as it severely limits the scope for inducing endogenous beneficial changes in culture.

Discussion

As cultures mix with increasing frequency, conflicts involving irreconcilable viewpoints are inevitable[3]. Ultimately, social and economic forces will almost certainly lead some traditions in some cultures to give way. That said, policy makers can resolve conflicts in ways that respect cultural differences, relatively speaking, to the extent that change is endogenous to the target population. This principle captures much of the policy appeal of tipping points and path-dependent changes in culture. The policy maker does not necessarily need draconian measures like criminalising female genital cutting[57,58]. Rather, the policy maker needs an effective intervention, delimited in time and space, to place the target population on a new cultural path consistent with policy objectives. The policy maker, in effect, wants to help people help themselves.

The appeal is clear, but the link between the psychology of social decision making and cultural change is varied and elaborate. Ample empirical evidence shows that decision makers are subject to positive social influence[27,38,59], but this is not necessarily useful to the policy maker. For example, when deciding whether to cut, families want to follow the trend around them to some extent[29,31,32]. This positive social effect at the family level, however, only comes into its own as a policy lever if it generates scope for spillovers at the aggregate level. Importantly, when positive social influence mixes with run-of-the-mill heterogeneity, spillovers can range from the utterly trivial to the truly spectacular. For this reason, estimating the scope for recruiting endogenous social mechanisms to advance behaviour change is crucial. Table 1 summarises the mechanisms examined and their effects. Here we turn to four associated principles to consider when aiming to trigger endogenous change.

First, the distribution of behaviours can provide the policy maker with clues about social influence and spillovers. A target population initially comprised of a locally heterogeneous mix, with many agents choosing the harmful option and many choosing some alternative[12,30,31], may not be a population suitable for generating spillovers. Local heterogeneity suggests that people are routinely mixing with others who think and behave differently. If endogenous social influence was going to send the population off in one direction or

another, it very well may have done so by the time the policy maker first arrives on the scene.

In contrast, if the harmful practice is locally pervasive, at least two possibilities hold. Either the target population is in the harmful equilibrium of a path-dependent process, or an intrinsic preference for the harmful behaviour ensures that the only stable equilibrium is harmful. In the former case, the policy maker needs a delimited intervention to push the population across the tipping point. Whatever the intervention may be, the hope is that social influence can help the policy maker avoid a protracted and potentially heavy-handed campaign. In the latter case, the policy maker must find out why members of the target population prefer a seemingly harmful option and how she can effectively change this preference. Moreover, she must do so without risking backlash by activating any tendency for people to view their cultural identities as under threat[3,12,54]. In situations of this kind, path dependence and multiple equilibria are irrelevant.

Second, coordination incentives with heterogeneous preferences support a diverse array of outcomes, but targeting a random sample of agents should generate spillovers more reliably than a biased sample. In general, people will differ in terms of their preferences, their likely response to a policy maker's intervention, and their networks. When people are heterogeneous in these ways, spillovers can range from the utterly trivial to the truly spectacular, and this remains true even if everyone has a shared interest in coordinating. Simply knowing that coordination incentives are widespread is not enough, and the risk is that the policy maker assumes spillovers can help her promote change when in reality they cannot.

Interestingly, when coordination incentives are pervasive in a heterogeneous population, our results suggest that a randomly selected target offers the most robust approach to sparking spillovers. A random target also has an important practical advantage. Namely, even if the policy maker does not have the option to survey the distribution of preferences in the target population ex ante, she can still target a random sample. She simply needs to avoid selection bias. Of particular significance, an amenable target is probably the worst strategy if the goal is to generate spillovers. An amenable target essentially takes the easiest possible task for the intervention and reserves the hardest possible task for endogenous social mechanisms. In spite of this, programmes promoting the abandonment of cutting have a tradition of working with individuals and communities favourable towards abandonment[3,11,12].

Third, when a concern for group identity is helping to sustain a harmful tradition, the policy maker should attempt to weaken the link between tradition and identity. When a group relies on the harmful practice to construct and maintain a distinct cultural identity, social influence is parochial. This means a focal decision maker wants to be like ingroup members but unlike outgroup members. The link between tradition and identity adds intrinsic value to the harmful practice for a subset of the population. This value, in turn, works against spillovers under nearly all conditions. In such a situation, the policy maker should consider strategies to curtail the effects of parochial social influence.

As one approach, she can accept that distinct groups within the target population want to construct oppositional identities, but she can try to transfer this dynamic to some new decision-making domain with reduced potential for harm. Candidate domains are likely to be highly idiosyncratic and specific to the groups in question. Generically, however, the hope is that the alternative choice domain will provide a new and relatively innocuous basis for groups to distinguish themselves. To the extent that the harmful practice is a traditional component of group identity, this strategy may be difficult. Moreover, as the example from the Meru district of Kenya illustrates, an especially challenging scenario occurs when the target population takes the policy maker and her foreign constituency as the relevant outgroup[3,12]. Transferring concerns about cultural identity to some other decision-making domain may be impossible in such cases precisely because the policy maker herself has established the harmful practice as an issue of central importance.

Alternatively, the policy maker can attempt to convince agents they do not need to construct oppositional identities. In this case, the policy maker faces two distinct but linked challenges. She must ensure that all groups abandon their parochial stance, and having done so she must specifically convince the groups with the harmful tradition to join the others. For the various reasons outlined above, mobilising social effects to induce endogenous change can be a formidable challenge in its own right. With parochialism in the mix, the challenge is doubly serious because outgroup anti-conformity typically compromises the potential for beneficial spillovers, and when strong it places significant constraints on behaviour change among agents not directly exposed to the intervention. For these reasons, if a harmful practice like cutting is an important component of group identity, the policy maker cannot rely on endogenous change without also addressing the fact that agents use the harmful practice to define themselves culturally.

Fourth, pre-existing preferences are pivotal in terms of what the policy maker can expect from social influence. Specifically, we have shown that spillovers hinge critically on the initial distribution of preferences in the target population and the details of how an intervention transforms this distribution. That said, in the analyses above an exogenous intervention reduces thresholds to zero for some families. Subsequent spillovers are limited to changes in behaviour; endogenous changes in preferences do not occur. However, if social influence were also to support endogeneous preference change, the spread of preferences favouring abandonment could be a powerful mechanism in its own right[12]. For this to be true, the cultural evolution of preferences would have to be somehow congruent with but different from the cultural evolution of behaviour.

Specifically, if preferences and behaviour always change in unison, preference change is redundant from a policy perspective. Imagine, however, that preferences and choices both respond to changes in the frequency of families who do not cut, but preference change and behaviour change are somehow different. This would introduce the possibility that some families first change their preferences in favour of abandonment, only to abandon cutting later when the population crosses the new reduced threshold values for these families.

This additional path to behaviour change could improve conditions for spillovers, but outcomes should again depend on the precise details of how an intervention interacts with the pre-existing characteristics of the target population.

Analogous complexities should apply to interventions that convince opinion leaders to abandon a harmful practice. Fieldwork on cutting[3] suggests that, when prominent members of an ingroup or outgroup voice their support for abandonment, the effect can go either way. If people view the call for abandonment as legitimate and genuine, and if they take the opinion leader as a valid normative model, the contribution to abandonment can be significant. However, if people feel the call for abandonment represents pandering to outside forces, or if the opinion leader does not provide a valid normative example, the effect can actually reinforce the commitment to cutting, as in the Meru example above.

Aside from the possible significance of endogenous preference change and influential individuals, our results have specific limitations to bear in mind. To motivate our assumptions, we have turned to empirical studies on cutting with data from Senegal to Sudan, from Switzerland to Kenya. This diversity mirrors the extensive geographic span of cutting. Because of this diversity, the mix of social mechanisms that supports cutting, and that could potentially contribute to cutting's decline, may vary considerably from one location to another. Consequently, our focus on mechanisms rooted in coordination and conformity may be more or less relevant in one location versus another. Indeed, aside from individual heterogeneity in conformist tendencies within a culture, recent research has also shown variation in generic tendencies to conform across cultures[60]. We have emphasized coordination and conformity in part because they have been highly influential in policy discussions. Coordination and conformity can, under the right circumstances, stabilize multiple outcomes, at least one of which the policy maker disfavours and at least one of which the policy maker favours. Aside from coordination and conformity, however, a vast number of other social mechanisms and psychological biases may operate when people influence each other and learn from each other. For many mechanisms and biases, the aggregate consequences remain unexplored.

All in all, conformity and coordination incentives at the individual level can translate into aggregate behavioural dynamics in diverse and subtle ways. Mundane forms of heterogeneity are crucial. If people vary in terms of whom they know and how they react to these people, conformity and coordination can support the abandonment of a harmful practice via spillovers, they can hinder change, or they can do nothing at all. At present, we have considerable evidence that conformity and coordination affect the choices families make about cutting[31,32]. We also, however, have considerable evidence for local heterogeneity in attitudes and practices related to cutting[30,31]. This is a combination that should lead policy makers to hope that spillovers are possible but worry that they may not be reliable, and indeed recent research finds little evidence for spillovers[12,58].

Spillovers are attractive in a world where cultures increasingly mix, and thus cultural conflicts are likely to be common. Any strategy that can manage these conflicts without one culture steamrolling another is a welcome strategy. Nonetheless, exploiting the potential of spillovers will require a concentrated effort to measure and evaluate how social influence and everyday heterogeneity combine to support the endogenous abandonment of harmful traditions like female genital cutting.

Methods

The characteristic normalised spillover

To standardise endogenous outcomes regardless of the size of an intervention, we normalise spillovers by accounting for the scale of the policy maker's intervention. Consider a set of independent populations, S. S could be a set of disjoint communities for empirical work, but here we will often treat S as a set of 200 independent populations whose cultural evolutionary dynamics we have simulated (Supplementary Information). For any $s \in S$, let $\hat{q}_s$ be the long-run proportion of agents who do not cut after endogenous changes have run their course in the wake of the intervention, and the population has re-stabilised. We define the characteristic normalised spillover,

$$\Theta_s = \max\left\{0, (1/|S|)\sum_{s\in S}\frac{[\hat{q}_s > \phi](\hat{q}_s - \phi)}{1-\phi} + \frac{[\phi \geq \hat{q}_s](\hat{q}_s - \phi)}{\phi}\right\}, \tag{1}$$

where $|\cdot|$ is set cardinality, and $[\cdot]$ are Iverson brackets. In effect, the final outcome in each population is normalised as a number between -1 and 1. Negative outcomes (i.e. $(\hat{q}_s - \phi)/\phi \in [-1, 0)$) occur, for example, if the proportion of decision makers choosing not to cut actually decreases after the intervention. The extreme value of -1 would occur if all agents initially cut, some agents abandon cutting in response to the intervention, but they return to cutting after the intervention ends. A value of 0 occurs if the proportion who do not cut after the intervention stabilises on the intervention size (i.e. $\hat{q}_s = \phi$). Such a value arises, for example, if all agents initially cut, targeted agents abandon cutting in response to the intervention, and no one changes their choices after the intervention. Positive outcomes occur (i.e. $(\hat{q}_s - \phi)/(1-\phi) \in (0, 1]$) if the proportion of decision makers choosing not to cut increases due to spillovers after the intervention ends. The extreme value of 1 would occur if all agents initially cut, some abandon cutting because of the intervention, and everyone else follows after the intervention ends. For the characteristic normalised spillover, we average normalised outcomes and take either this average or zero according to which is larger.

Spillovers under alternative intervention targets and heterogeneous responses to the intervention

For simulations (Supplementary Information), we consider 121 different initial threshold distributions (Supplementary Information, Supplementary Fig. 15), five different values of ϕ ranging from 0.1 to 0.5, and the three different intervention strategies (amenable, random, resistant). For agent i, with pre-intervention threshold q_i, we assume the agent responds to the intervention with probability $h(q_i) = 1 - cq_i$, where $c \in \{0, 0.5, 1\}$.

We modelled homophily in three different ways (Supplementary Information). In all cases, homophily has the following generic feature. As the pre-intervention thresholds for two agents become increasingly far apart, the probability the two agents are connected to each other declines. We consider various degrees of homophily, from extremely weak to strong. As homophily becomes increasingly strong, the population becomes increasingly fragmented. As a result, agents only know about the choices of some people in the population.

Importantly, the detrimental effects of homophily do not occur simply because homophily fragments the population. As a kind of benchmark, we repeated the exercise shown in Figs. 5-6, but we used random networks with linkage probabilities ranging from zero to one. Spillovers tend to be larger under random networks than under homophilous networks. Moreover, the detrimental effects of fragmentation typically only appear, quite suddenly, as the linkage probability gets close to zero (Supplementary Information, Supplementary Fig. 25-26). This is very different from the broadly detrimental effects of homophily (Figs. 5-6).

Spillovers when harmful traditions are tied to group identity

Assume the population consists of two groups (Supplementary Information). At a given point in time, each family evaluates its current practice, whether cutting or not cutting, and chooses either to retain its current practice or change. When doing so, a family can either respond to the ingroup distribution of choices or to the outgroup distribution. Specifically, use i to index family and j to index group. Let the choice of i in j in period t be a Bernoulli random variable, Y_{ijt}, such that $Y_{ijt} = 1$ indicates not cutting, and $Y_{ijt} = 0$ indicates cutting. Let X_{ijt} be a Bernoulli random variable such that $X_{ijt} = 1$ indicates the family responds to the ingroup distribution of choices, while $X_{ijt} = 0$ indicates a response to the outgroup. Let q_{jt} indicate the proportion of families in j not cutting in t.

With probability $\beta \in [0.5, 1]$, a family bases its current choice on the ingroup distribution of behaviours,

i.e. $P(X_{ijt} = 1) = \beta$. When doing so, families tend to conform to the majority choice,

$$P(Y_{ij(t+1)} = 1 \,|\, X_{ij(t+1)} = 1) = a_{ij} + \frac{(b_{ij} - a_{ij})q_{jt}^{\gamma_{ij}}}{q_{jt}^{\gamma_{ij}} + (1 - q_{jt})^{\gamma_{ij}}}, \tag{2}$$

where $a_{ij}, b_{ij} \in [0, 1]$, $\bar{a}_j < \bar{b}_j$, $\gamma_{ij} \geq 0$, and $\bar{\gamma}_j > 1$. Larger values of $\bar{\gamma}_j$ yield, given a response to the ingroup, relatively strong ingroup conformity. Because $\bar{a}_j < \bar{b}_j$ and $\bar{\gamma}_j > 1$, ingroup responses tend to be positively sloped sigmoidal functions that homogenise choices within groups, whether cutting or not cutting. If this ingroup conformity was the only relevant mechanism (i.e. $\beta = 1$), the result would be multiple equilibria, tipping points, and ample scope for endogenous spillovers.

With probability $1 - \beta$, a family bases its current choice on the outgroup distribution of behaviours, i.e. $P(X_{ijt} = 0) = 1 - \beta$. When doing so, families typically take the majority choice in the outgroup as an indication of what not to do,

$$P(Y_{ij(t+1)} = 1 \,|\, X_{ij(t+1)} = 0) = c_{ij} + \frac{(d_{ij} - c_{ij})q_{j't}^{\mu_{ij}}}{q_{j't}^{\mu_{ij}} + (1 - q_{j't})^{\mu_{ij}}}, \tag{3}$$

where $c_{ij}, d_{ij} \in [0, 1]$, $\bar{c}_j > \bar{d}_j$, $\mu_{ij} \geq 0$, and $\bar{\mu}_j > 1$. In addition, $j \neq j'$, which simply means (3) is an outgroup response. Larger values of $\bar{\mu}_j$ yield, given a response to the outgroup, relatively strong outgroup anti-conformity. This outgroup anti-conformity represents an effort to establish and maintain distinct group identities by using, with probability $1 - \beta$, the outgroup majority as an example of how not to behave. This follows from $\bar{c}_j > \bar{d}_j$ and $\bar{\mu}_j > 1$, which ensures that out-group responses tend to be negatively sloped sigmoidal functions.

The intervention targets a proportion ϕ. All targeted agents abandon unconditionally (i.e. $a_{ij}, b_{ij}, c_{ij}, d_{ij} = 1$), and thus ϕ is a lower bound on the proportion not cutting after the intervention. When β is close to one, the two groups are approximately independent of each other, and no one cares much if the two groups have traditions that are different or similar. As β approaches 0.5, however, group identity becomes increasingly important, and when $\beta = 0.5$ ingroup conformity and outgroup anti-conformity are equally important.

With this structure in place, we model cultural evolutionary dynamics in two ways (Supplementary Information). First, we assume an infinitely large population of agents who are homogeneous. This simplification allows us to treat cultural evolutionary dynamics as deterministic, and we use a mix of analytical and graphical techniques to analyse the model. Second, we assume a finite population of heterogeneous agents. We consider a variety of distributions for controlling heterogeneity, but given the question of interest we focus on cases in which the expected response to the ingroup is to conform to the majority and the expected response to the outgroup is to anti-conform. A large empirical literature on cultural transmission shows that people use ingroup conformist strategies, and in some cases outgroup anti-conformist strategies[61], but these strate-

gies tend to be highly variable[36–39,62,63]. For this reason, we simulate cultural evolutionary dynamics under a wide array of parameter values for controlling these strategies, and we focus attention on results robust to this variation. For a given combination of parameter values, we simulate 200 independent populations and use the results to estimate the characteristic normalised spillover.

To identify the conditions depicted in Fig. 7, we ran an initial set of simulations to find the strength of ingroup conformity ($\bar{\gamma}_j$) that leads to the largest spillovers when group identity concerns are absent ($\beta = 1$). We then chose a range of $\bar{\gamma}_j$ and $\bar{\mu}_j$ values in the vicinity of the spillover-maximizing $\bar{\gamma}_j$ value. We varied $\bar{\gamma}_j$ and $\bar{\mu}_j$ values independently, and for any combination of values we simulated cultural evolution for β from 1.0 down to 0.5. Fig. 7 summarises results from this exercise for intervention sizes from $\phi = 0.1$ to $\phi = 0.5$.

Code availability

Code is available as Supplementary Software with related details in the Supplementary Information and the Supplementary Software Guide.

References

[1] Nyborg, K. *et al.* Social norms as solutions. *Science* **354**, 42–43 (2016).

[2] Shell-Duncan, B. & Hernlund, Y. Female "circumcision" in Africa: dimensions of the practice and debates. In Shell-Duncan, B. & Hernlund, Y. (eds.) *Female "Circumcision" in Africa: Culture, Controversy, and Change*, 1–40 (Boulder, CO: Lynne Rienner, 2000).

[3] Cloward, K. *When Norms Collide: Local Responses to Activism Against Female Genital Mutilation and Early Marriage* (Oxford University Press, 2016).

[4] Shell-Duncan, B. From health to human rights: female genital cutting and the politics of intervention. *American Anthropologist* **110**, 225–236 (2008).

[5] Richerson, P. J. & Boyd, R. *Not By Genes Alone: How Culture Transformed the Evolutionary Process* (Chicago: University of Chicago Press, 2005).

[6] Dolan, P. *et al.* Influencing behaviour: the mindspace way. *Journal of Economic Psychology* **33**, 264–277 (2012).

[7] World Bank Group. *Mind, Society, and Behavior: World Development Report 2015* (Washington DC: The World Bank, 2015).

[8] Bicchieri, C. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms* (Oxford University Press, 2016).

[9] Shell-Duncan, B., Wander, K., Hernlund, Y. & Moreau, A. Dynamics of change in the practice of female genital cutting in Senegambia. *Social Science & Medicine* **73**, 1275–1283 (2011).

[10] UNFPA-UNICEF. Joint evaluation of the UNFPA-UNICEF joint programme on female genital mutilation/cutting: Accelerating change. Preprint at `https://www.unfpa.org/admin-resource/unfpa-unicef-joint-evaluation-unfpa-unicef-joint-programme-female-genital` (2013).

[11] Mackie, G., Moneti, F., Shakya, H. & Denny, E. What are social norms? how are they measured. Preprint at `https://www.unicef.org/protection/files/4_09_30_Whole_What_are_Social_Norms.pdf` (2015).

[12] Platteau, J.-P., Camilotti, G. & Auriol, E. Eradicating women-hurting customs. In Anderson, S., Beaman, L. & Platteau, J. (eds.) *Towards Gender Equity in Development*, 319–356 (Oxford University Press, 2018).

[13] Malhotra, A. and Warner, A. and McGonagle, A. and Lee-Rife, S.. Solutions to end child marriage: what the evidence shows. Preprint at `https://www.icrw.org/wp-content/uploads/2016/10/Solutions-to-End-Child-Marriage.pdf` (2011).

[14] Lee-Rife, S., Malhotra, A., Warner, A. & Glinski, A. M. What works to prevent child marriage: a review of the evidence. *Studies in Family Planning* **43**, 287–303 (2012).

[15] Bicchieri, C., Jiang, T. & Lindemans, J. W. A social norms perspective on child marriage: the general framework. Preprint at `http://repository.upenn.edu/pennsong/13/` (2014).

[16] Shakya, H. B., Christakis, N. A. & Fowler, J. H. Social network predictors of latrine ownership. *Social Science & Medicine* **125**, 129–138 (2015).

[17] World Health Organization. Changing cultural and social norms that support violence. Preprint at `https://apps.who.int/iris/handle/10665/44147` (2009).

[18] Christakis, N. A. & Fowler, J. H. The collective dynamics of smoking in a large social network. *New England Journal of Medicine* **358**, 2249–2258 (2008).

[19] Mackie, G. Ending footbinding and infibulation: a convention account. *American Sociological Review* **61**, 999–1017 (1996).

[20] Prentice, D. A. & Miller, D. T. Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology* **64**, 243–256 (1993).

[21] Young, H. P. The evolution of social norms. *Annual Review of Economics* **7**, 359–387 (2015).

[22] Christakis, N. A. & Fowler, J. H. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine* **2007**, 370–379 (2007).

[23] Paluck, E. L., Shepherd, H. & Aronow, P. M. Changing climates of conflict: a social network experiment in 56 schools. *Proceedings of the National Academy of Sciences* **113**, 566–571 (2016).

[24] Allcott, H. Social norms and energy conservation. *Journal of Public Economics* **95**, 1082–1095 (2011).

[25] Hallsworth, M., List, J. A., Metcalfe, R. D. & Vlaev, I. The behavioralist as tax collector: using natural field experiments to enhance tax compliance. *Journal of Public Economics* **148**, 14–31 (2017).

[26] Castilla-Rho, J. C., Rojas, R., Andersen, M. S., Holley, C. & Mariethoz, G. Social tipping points in global groundwater management. *Nature Human Behaviour* **1**, 640 (2017).

[27] Koch, C. M. & Nax, H. H. Groundwater: testing the behavioral foundations of commons problems. Preprint at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3075935` (2017).

[28] World Health Organization. Female genital mutilation. `http://www.who.int/news-room/fact-sheets/detail/female-genital-mutilation` (2018).

[29] Hayford, S. R. Conformity and change: community effects on female genital cutting in Kenya. *Journal of Health and Social Behavior* **46**, 121–140 (2005).

[30] Bellemare, M. F., Novak, L. & Steinmetz, T. L. All in the family: explaining the persistence of female genital cutting in West Africa. *Journal of Development Economics* **116**, 252 – 265 (2015).

[31] Efferson, C., Vogt, S., Elhadi, A., Ahmed, H. E. F. & Fehr, E. Female genital cutting is not a social coordination norm. *Science* **349**, 1446–1447 (2015).

[32] Howard, J. A. & Gibson, M. A. Frequency-dependent female genital cutting behaviour confers evolutionary fitness benefits. *Nature Ecology & Evolution* **1**, 0049 (2017).

[33] Vogt, S., Zaid, N. A. M., Ahmed, H. E. F., Fehr, E. & Efferson, C. Changing cultural attitudes towards female genital cutting. *Nature* **538**, 506–509 (2016).

[34] De Cao, E. & Lutz, C. Sensitive survey questions: measuring attitudes regarding female genital cutting through a list experiment. *Oxford Bulletin of Economics and Statistics* **80**, 871–892 (2018).

[35] Gibson, M. A., Gurmu, E., Cobo, B., Rueda, M. M. & Scott, I. M. Indirect questioning method reveals hidden support for female genital cutting in South Central Ethiopia. *PloS ONE* **13**, e0193985 (2018).

[36] Efferson, C., Lalive, R., Richerson, P. J., McElreath, R. & Lubell, M. Conformists and mavericks: the empirics of frequency-dependent cultural transmission. *Evolution and Human Behavior* **29**, 56–65 (2008).

[37] Morgan, T. J. H., Rendell, L. E., Ehn, M., Hoppitt, W. & Laland, K. N. The evolutionary basis of human social learning. *Proceedings of the Royal Society B* **279**, 653–662 (2012).

[38] Muthukrishna, M., Morgan, T. J. & Henrich, J. The when and who of social learning and conformist transmission. *Evolution and Human Behavior.* **37**, 10–20 (2016).

[39] Efferson, C. & Vogt, S. Behavioural homogenization with spillovers in a normative domain. *Proc. R. Soc. B* **285**, 20180492 (2018).

[40] Shell-Duncan, B. & Hernlund, Y. Are there "stages of change" in the practice of female genital cutting? qualitative research findings from Senegal and the Gambia. *African Journal of Reproductive Health* **10**, 57–71 (2006).

[41] Howard, J. A. & Gibson, M. A. Is there a link between paternity concern and female genital cutting in West Africa? *Evolution and Human Behavior* **40**, 1–11 (2019).

[42] Granovetter, M. Threshold models of collective behavior. *American Journal of Sociology* **83**, 1420–1443 (1978).

[43] Watts, D. J. & Dodds, P. Threshold models of social influence. *The Oxford Handbook of Analytical Sociology* 475–497 (2009).

[44] Young, H. P. Innovation diffusion in heterogeneous populations: contagion, social influence, and social learning. *American Economic Review* **99**, 1899–1924 (2009).

[45] Vogt, S., Efferson, C. & Fehr, E. The risk of female genital cutting in europe: comparing immigrant attitudes toward uncut girls with attitudes in a practicing country. *SSM-Population Health* **3**, 283–293 (2017).

[46] Schelling, T. C. Hockey helmets, concealed weapons, and daylight saving: a study of binary choices with externalities. *Journal of Conflict Resolution* **17**, 381–428 (1973).

[47] Xie, J. *et al.* Social consensus through the influence of committed minorities. *Physical Review E* **84**, 011130 (2011).

[48] Centola, D., Becker, J., Brackbill, D. & Baronchelli, A. Experimental evidence for tipping points in social convention. *Science* **360**, 1116–1119 (2018).

[49] Baronchelli, A., Felici, M., Loreto, V., Caglioti, E. & Steels, L. Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics: Theory and Experiment* **2006**, P06014 (2006).

[50] McPherson, M., Smith-Lovin, L. & Cook, J. M. Birds of a feather: homophily in social networks. *Annual Review of Sociology* **27**, 415–444 (2001).

[51] Jackson, M. O. & López-Pintado, D. Diffusion and contagion in networks with heterogeneous agents and homophily. *Network Science* **1**, 49–67 (2013).

[52] Young, H. P. The dynamics of social innovation. *Proceedings of the National Academy of Sciences* **108**, 21285–21291 (2011).

[53] Lu, Q., Korniss, G. & Szymanski, B. K. The naming game in social networks: community formation and consensus engineering. *Journal of Economic Interaction and Coordination* **4**, 221 (2009).

[54] Thomas, L. "Ngaitana (I Will Circumcise Myself)": lessons from colonial campaigns to ban excision in Meru, Kenya. In Shell-Duncan, B. & Hernlund, Y. (eds.) *Female "Circumcision" in Africa: Culture, Controversy, and Change*, 129–150 (Boulder, CO: Lynne Rienner, 2000).

[55] Gruenbaum, E. *The Female Circumcision Controversy: An Anthropological Perspective* (Philadelphia: University of Pennsylvania Press, 2001).

[56] Goodman, R. & Jinks, D. *Socializing States: Promoting Human Rights through International Law* (Oxford University Press, 2013).

[57] Shell-Duncan, B., Wander, K., Hernlund, Y. & Moreau, A. Legislating change? responses to criminalizing female genital cutting in Senegal. *Law & Society Review* **47**, 803–835 (2013).

[58] Camilotti, G. Interventions to stop female genital cutting and the evolution of the custom: evidence on age at cutting in senegal. *Journal of African Economies* **25**, 133–158 (2016).

[59] Morgan, T. J. H. & Laland, K. N. The biological bases of conformity. *Frontiers in Neuroscience* **6** (2012).

[60] Molleman, L. & Gächter, S. Societal background influences social learning in cooperative decision making. *Evolution and Human Behavior* **39**, 547 – 555 (2018).

[61] Efferson, C., Lalive, R., Cacault, M. P. & Kistler, D. The evolution of facultative conformity based on similarity. *PLoS ONE* **11**, e0168551 (2016).

[62] Molleman, L., van den Berg, P. & Weissing, F. J. Consistent individual differences in human social learning strategies. *Nature Communications* **5**, 3570 (2014).

[63] Mesoudi, A., Chang, L., Dall, S. R. X. & Thornton, A. The evolution of individual and cultural variation in social learning. *Trends in Ecology & Evolution* **31**, 215–225 (2016).

[64] Boyd, R. & Richerson, P. J. *Culture and the Evolutionary Process* (Chicago: University of Chicago Press, 1985).

[65] Krumpal, I. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity* **47**, 2025–2047 (2013).

[66] Apicella, C. L., Marlowe, F. W., Fowler, J. H. & Christakis, N. A. Social networks and cooperation in hunter-gatherers. *Nature* **481**, 497 (2012).

[67] Migliano, A. *et al.* Characterization of hunter-gatherer networks and implications for cumulative culture. *Nature Human Behaviour* **1**, 0043 (2017).

[68] Centola, D. An experimental study of homophily in the adoption of health behavior. *Science* **334**, 1269–1272 (2011).

Acknowledgements

For valuable comments while developing this research, we thank James Walsh, as well as seminar participants at the Universities of Bern, Konstanz, Lausanne, Nottingham, and Zurich, the United Nations University in Maastricht, Harvard, and Oxford. CE and SV would also like to thank the Swiss National Science Foundation (Grant No. 100018_185417/1). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

CE designed, implemented, and analysed the models. SV surveyed the relevant policy literature. CE wrote the paper with input from SV and EF.

Competing interests

The authors declare no competing interests.

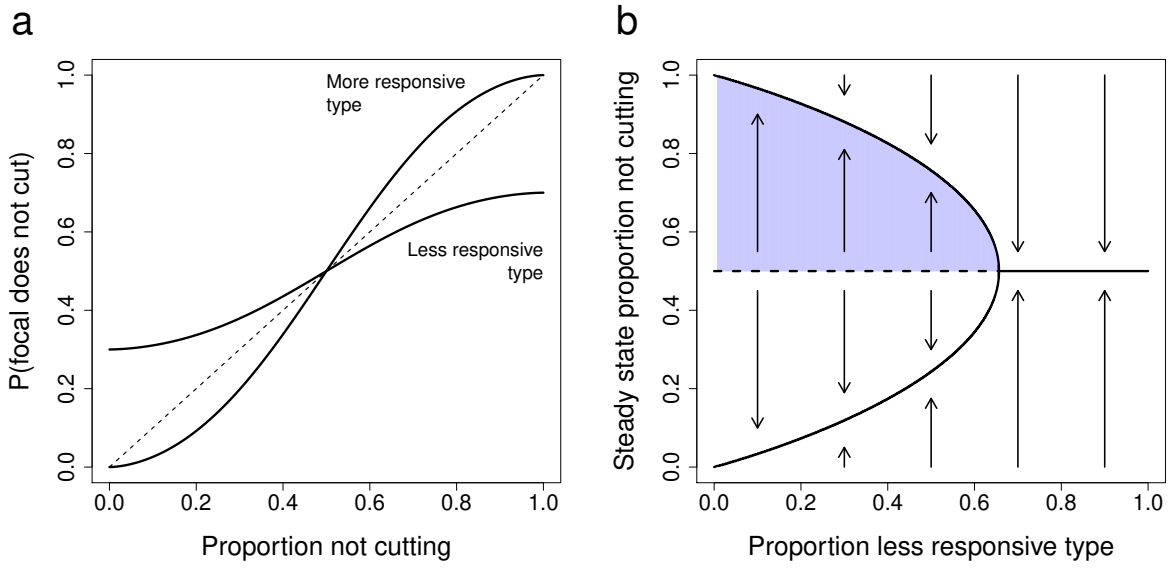

Figure 1: **Heterogeneity and spillovers. a**, Assume the target population consists of two types. A focal decision maker of either type is more likely to choose not cutting with increases in the proportion of decision makers who do not cut, and for either type this response exhibits the sigmoidal shape associated with path-dependent dynamics[59,64] and potential spillovers. The two types vary in terms of how responsive to social information they are[31,36], with one type more responsive than the other. **b**, Solid lines show stable steady states as a function of how common the less responsive type is, and the dashed line shows the unstable steady state. Arrows show the direction of cultural evolution. When two stable steady states exist, pushing the population from the lower stable steady state across the unstable steady state should induce a transition to the upper stable steady state, with the maximum possible spillover shown in purple. As the less responsive type becomes more common in the population, all stable steady states converge on the one steady state supported by this type. The potential for beneficial spillovers declines until it disappears altogether. More broadly, the structure of heterogeneity controls the extent to which the policy maker can rely on endogenous cultural change.

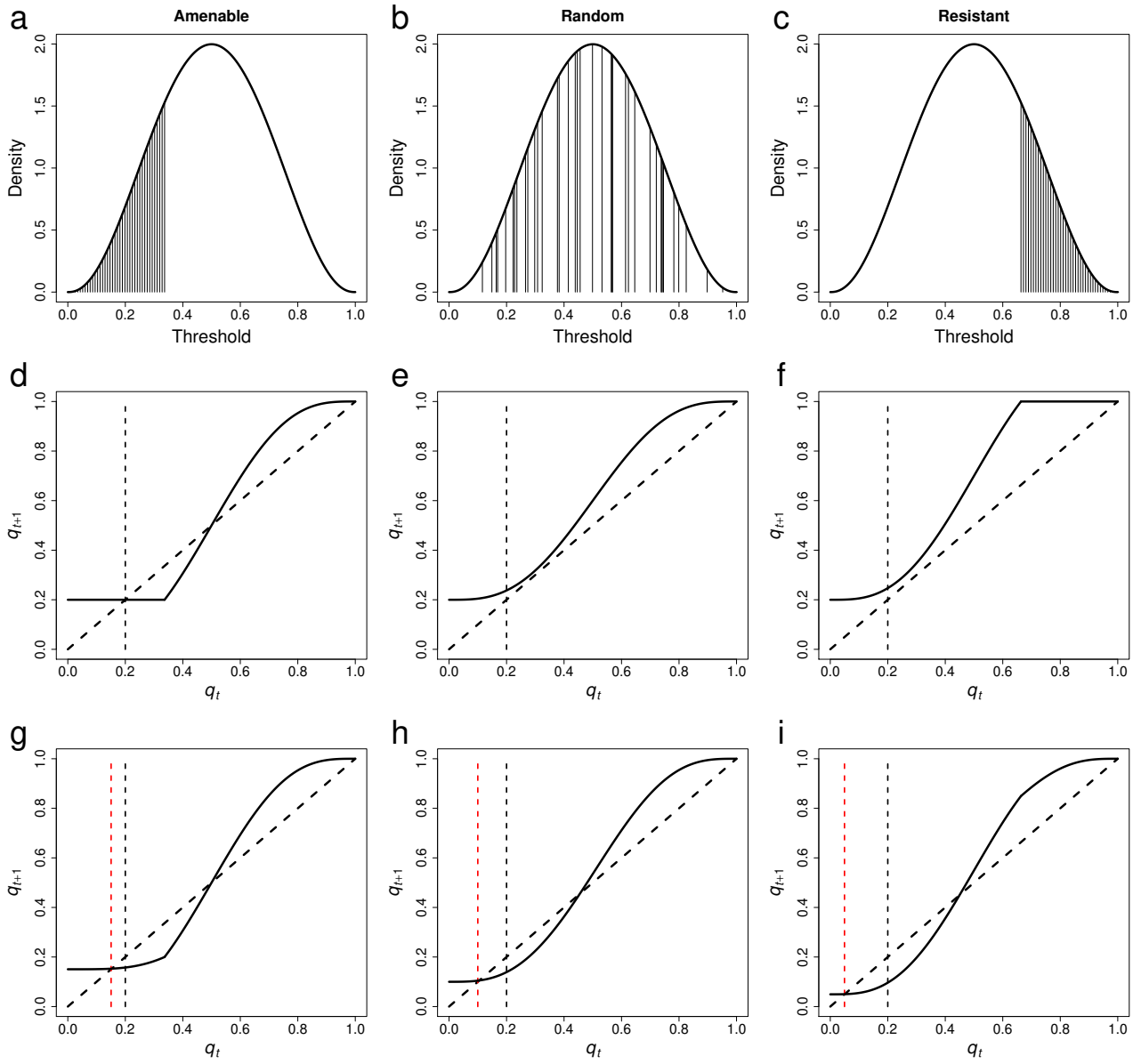

Figure 2: **Variation in intervention targets. a-c**, An example density function showing the distribution of threshold values before the intervention. **d-i**, Associated cumulative distribution functions after the intervention when everyone targeted responds (**d-f**) and when the probability of responding declines linearly with an agent's initial threshold value (**g-i**). Cumulative distribution functions specify cultural evolutionary dynamics, i.e. $q_{t+1} = F(q_t)$. $F(q_t) > q_t$ implies an increase in abandonment, $F(q_t) < q_t$ an increase in cutting. **a, d**, The intervention targets 20% of agents (**d**, vertical dashed line) with the lowest initial threshold values (**a**, shaded region). They abandon cutting unconditionally. 20% not cutting becomes a stable equilibrium as a result, and no spillovers occur. **b, e**, The intervention targets a random sample constituting 20% of agents. The cumulative distribution function shifts upwards, in this example clearing the 45° line. The only stable outcome is for everyone to abandon cutting, thus yielding the maximum possible spillover. **c, f**, A resistant target is similar, but the upwards shift is larger still. **g-i**, The proportion responding to the intervention (red vertical dashed lines) is less than the proportion targeted (black vertical dashed lines). This can eliminate the spillovers that would otherwise occur under random and resistant targets.

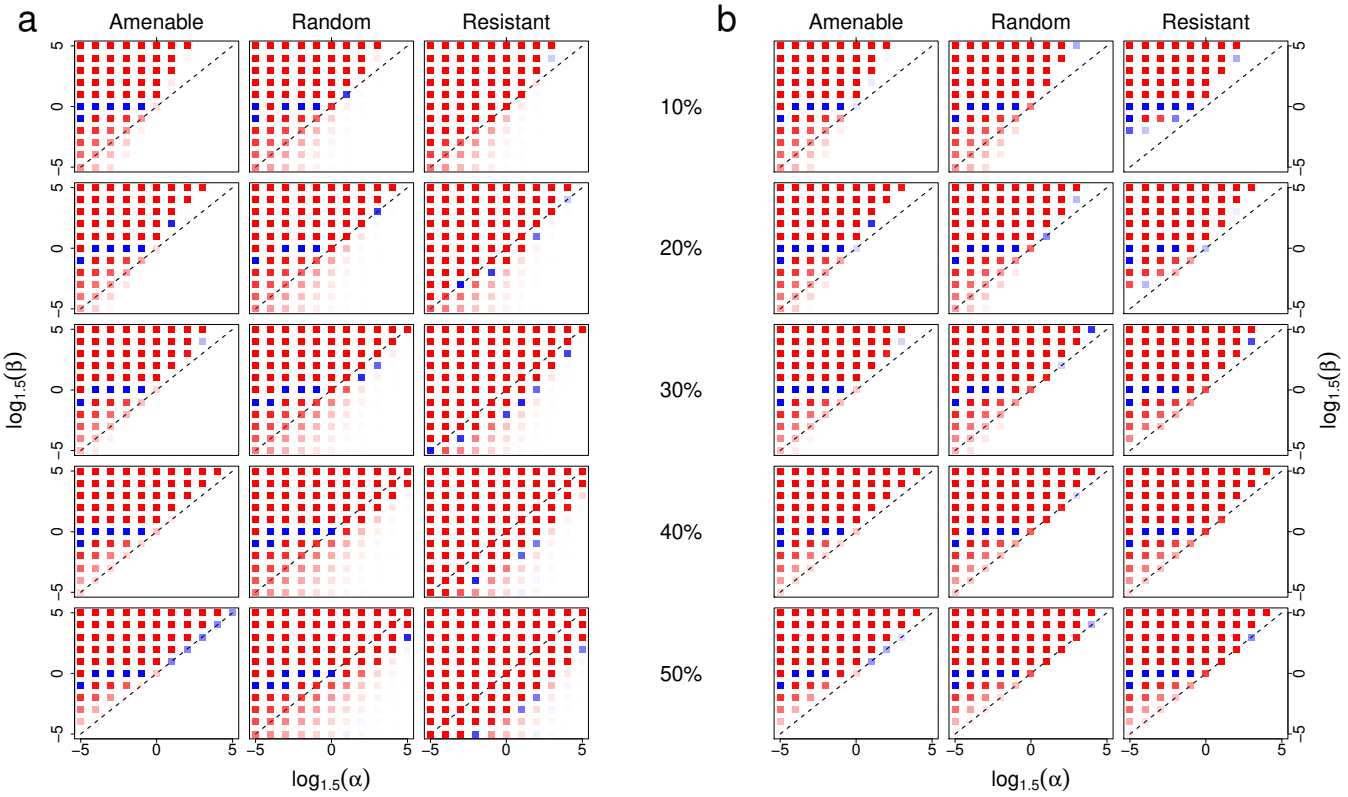

Figure 3: **The dominant effects of pre-existing preferences.** Squares vary in terms of which segment of the population the intervention targets (amenable to change, random, resistant to change) and the size of the intervention ($\{10\%, 20\%, \ldots, 50\%\}$). Each square shows simulation results under 121 different distributions of initial threshold values. Specifically, threshold values before the intervention are distributed Beta(α, β) (Supplementary Information, Supplementary Fig. 15). In the lower left quadrant of a square, distributions are bimodal. Above the 45° line, distributions are skewed right and thus favour abandonment. Below this line, distributions are skewed left and disfavour abandonment. Distributions along the 45° line are symmetric and neither favour nor disfavour abandonment. Colour intensity (red or blue) indicates the size of the characteristic normalised spillover, and thus white indicates no spillovers. Red signifies that spillovers are unimodally distributed across simulated populations of 500 agents each, and blue signifies a multimodal distribution. **a**, Fully connected networks in which everyone targeted by the intervention responds. Results show that the pre-existing threshold distribution has a dominant effect on spillovers. If this pre-existing distribution is approximately symmetric, the policy maker's choices also matter, and she does best by targeting agents resistant to change. **b**, Fully connected networks in which the probability of responding to the intervention decreases with an agent's initial threshold value. This heterogeneous response to the intervention destroys the advantages of a resistant target.

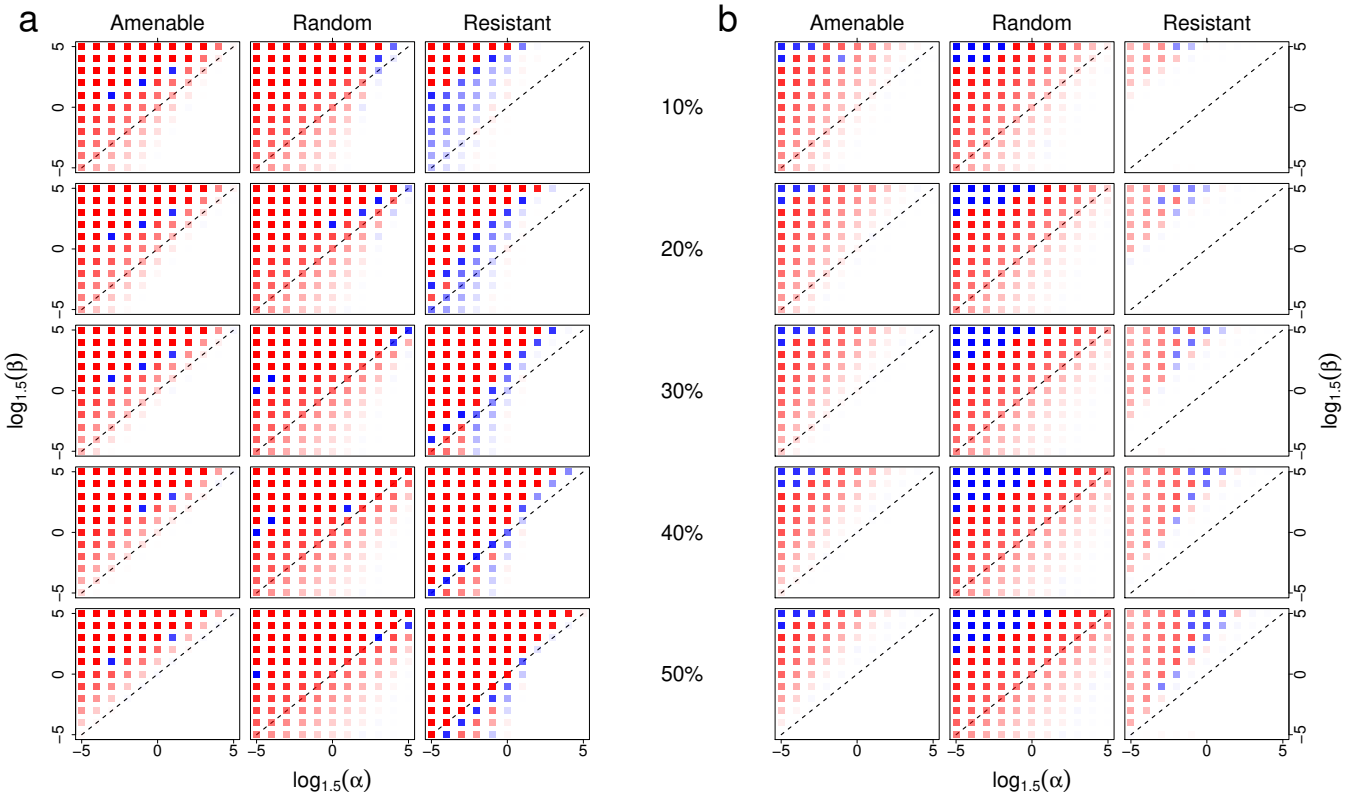

Figure 4: **Limited spillovers under homophily.** As in Fig. 3, each square shows results under 121 different distributions of initial threshold values. Specifically, threshold values before the intervention are distributed Beta(α, β) (Supplementary Information, Supplementary Fig. 15). Distributions above the 45° line favour abandonment, while distributions below this line do not. Squares vary in terms of both intervention targets (amenable, random, resistant) and the size of the intervention ($\{10\%, 20\%, \ldots, 50\%\}$). Colour intensity (red or blue) indicates characteristic normalised spillover values, with white meaning no spillovers. Red signifies a unimodal distribution of outcomes across simulated populations of 500 agents and blue a multimodal distribution. **a**, Moderately homophilous networks in which agents tend to link to others with similar initial threshold values, and the expected number of links per agent is 50.4. **b**, Strongly homophilous networks in which the expected number of links per agent is 5.5. Results indicate that a random target is the most robust approach to fostering spillovers under homophily.

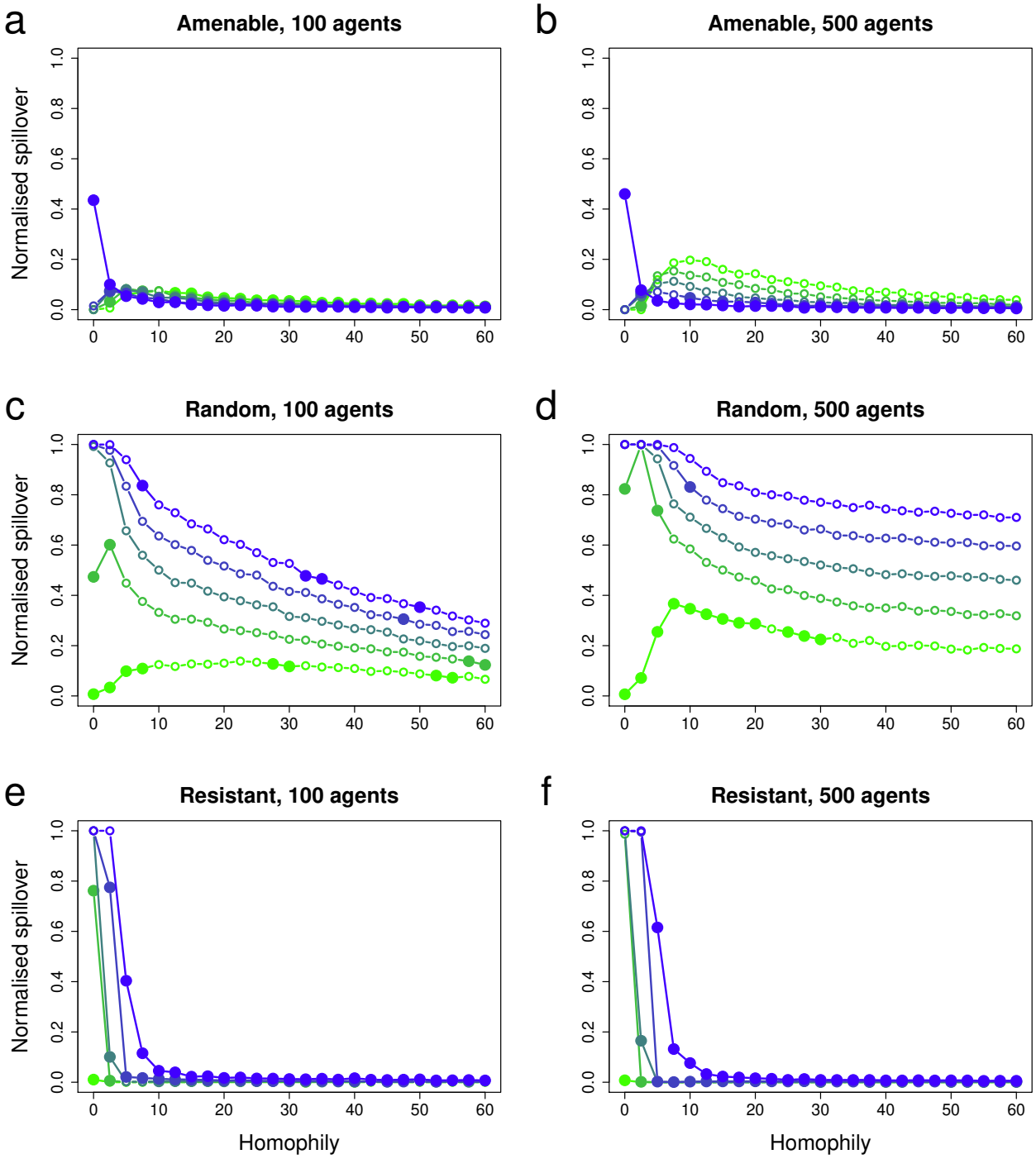

Figure 5: **The joint effects of selection bias and homophily.** Each panel shows characteristic normalised spillovers as a function of homophily (i.e. θ in Supplementary Information). Panels vary in terms of whether the intervention targets agents amenable for change (**a-b**), randomly selected agents (**c-d**), or agents resistant to change (**e-f**). The total number of agents is either 100 (**a, c, e**) or 500 (**b, d, f**). Interventions vary from 10% (green) to 50% (blue) of the population in increments of 10%. Targeted agents unconditionally adopt the policy maker's desired behaviour. Open circles indicate a unimodal distribution of outcomes across simulated populations. Closed circles indicate a multimodal distribution. Homophily tends to reduce beneficial spillovers. However, targeting a random sample of agents dramatically attenuates this effect compared to the biased selection associated with targeting amenable or resistant agents.

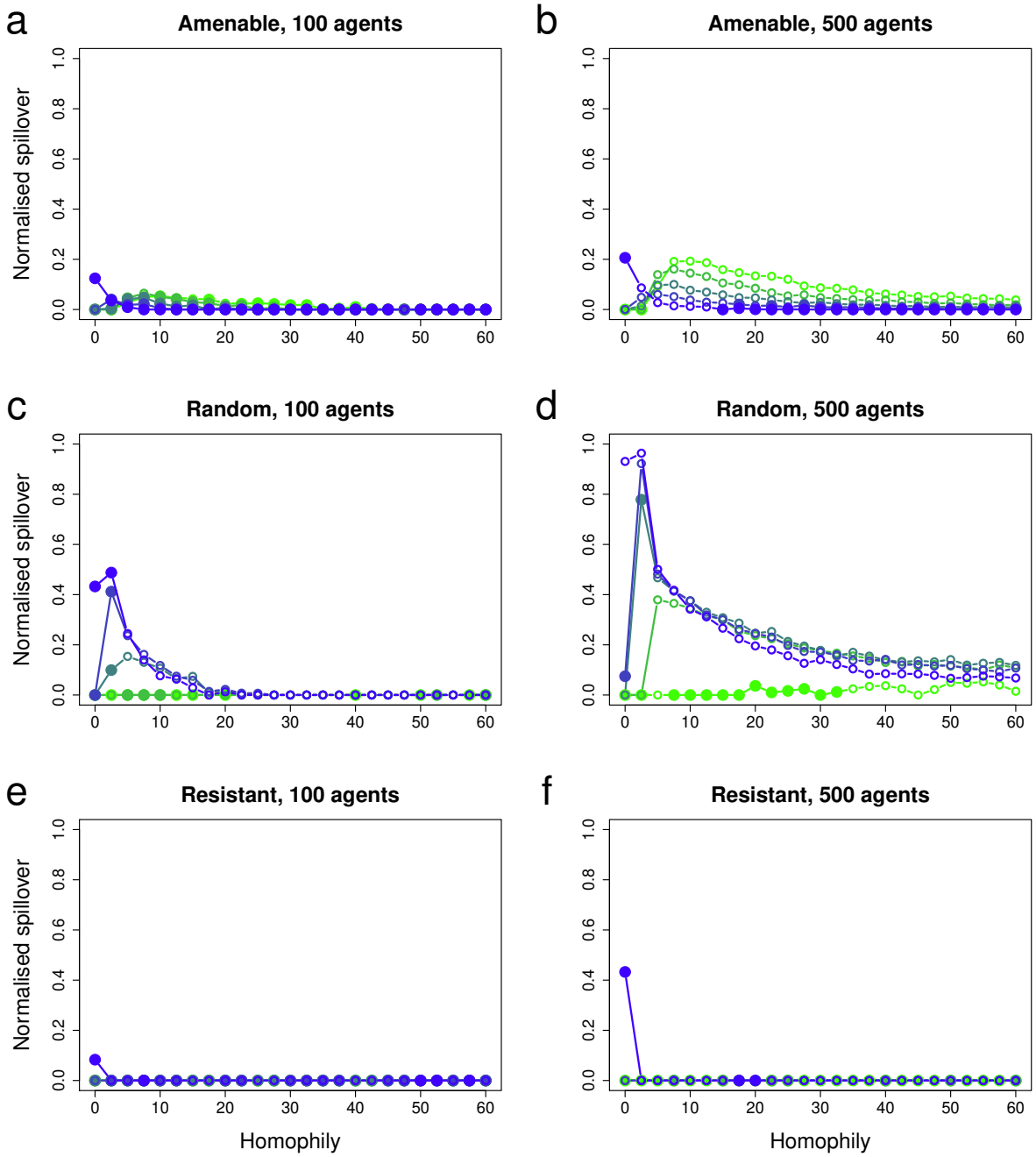

Figure 6: **Combining heterogeneous responses to the intervention with selection bias and homophily.** As in Fig. 5, each panel shows characteristic normalised spillovers as a function of homophily (i.e. θ in Supplementary Information). Panels vary in terms of intervention targets (**a-b**, amenable; **c-d**, random; **e-f**, resistant) and the total number of agents (**a, c, e**, 100; **b, d, f**, 500). Interventions vary from 10% (green) to 50% (blue) of the population in increments of 10%. Open circles indicate a unimodal distribution of outcomes across simulated populations. Closed circles indicate a multimodal distribution. Like Fig. 5, the results here show the joint effects of selection bias and homophily. Here we add the assumption that agents are decreasingly likely to respond to the intervention as initial threshold values increase, and thus only some agents targeted by the intervention adopt the policy maker's desired behaviour. This further reduces spillovers (cf. Fig. 5), but randomly selecting agents again moderates the effect.

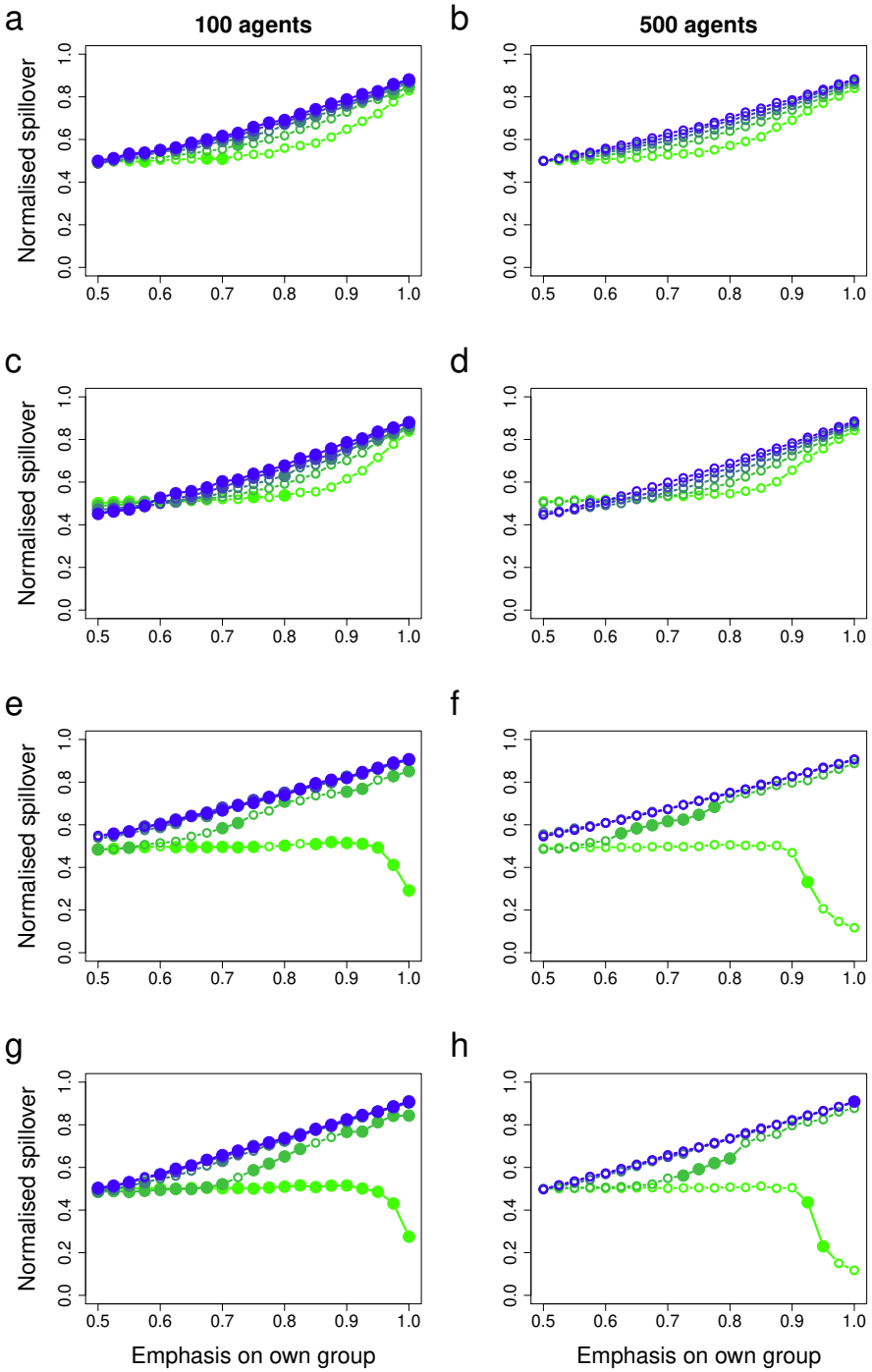

Figure 7: **Group identity as a drag on beneficial change.** As the emphasis on responding to one's ingroup (β, see Methods) declines, and by extension the emphasis on distinguishing one's own group from the outgroup ($1-\beta$) increases, spillovers decrease. Graphs show the characteristic normalised spillover in populations of 100 (**a, c, e, g**) and 500 agents (**b, d, f, h**). Interventions vary from 10% (green) to 50% (blue) of the population in increments of 10%. Targeted agents unconditionally adopt the policy maker's desired behaviour. Conditional on responding to the ingroup, ingroup conformity ($\bar{\gamma}_j$, see Methods) is relatively weak (**a-d**) or strong (**e-f**). Conditional on responding to the outgroup, outgroup anti-conformity ($\bar{\mu}_j$, see Methods) is also relatively weak (**a-b, e-f**) or strong (**c-d, g-h**). Open circles indicate a unimodal distribution of outcomes and closed circles multimodal.

Table 1: Key questions about underlying mechanisms with policy implications and relevant citations.

Key question about underlying mechanism	Implications	References
What are the consequences for spillovers when everyone responds to information about the frequencies of different behaviours, but people vary in their sensitivity to this information?	As less sensitive types become more common, the scope for spillovers declines (Fig. 1). Significant spillovers depend on a sizeable proportion of agents who are highly sensitive. The policy maker should thus identify how important conformity and coordination are to different types of decision maker in the target population.	Efferson et al. (2015)[31] use an anonymous self-administered questionnaire to show that, within Sudanese communities, people vary markedly in terms of the importance they attach to coordinated cutting practices. Communities also vary from each other in that the importance of coordination predominates in some but not in others. More generally, experimental designs like those in Muthukrishna et al. (2016)[38] and Efferson and Vogt (2018)[39] identify heterogeneity in conformist strategies and any associated tendency to homogenise attitudes and behaviour.
How do spillovers depend on pre-intervention heterogeneity in preferences related to the harmful practice and its beneficial alternative?	Spillovers tend to be reliably large if most decision makers are amenable to a change in favour of the beneficial behaviour and routinely absent if most decision makers are opposed to change (Fig. 3). Spillovers are most sensitive to policy choices if pre-existing preferences neither favour nor disfavour change. The policy maker needs reliable data, uncompromised by social desirability bias[65], on the distribution of attitudes and preferences in the target population.	Hayford (2005)[29], Bellemare et al. (2015)[30], Efferson et al. (2015)[31], Cloward (2016)[3], Vogt et al. (2017)[45], Gibson et al. (2018)[35], and Platteau et al. (2018)[12] present evidence for and discuss the implications of locally heterogeneous attitudes and preferences related to cutting. Efferson et al. (2015)[31], de Cao and Lutz[34], and Gibson et al. (2018)[35] present methods designed to reduce social desirability bias when identifying attitudes and preferences related to cutting.
How do spillovers vary, based on which segment of the population the policy maker targets, when either (i) the intervention is uniformly effective for all members of the population or (ii) the intervention has an effect that declines as agents become more resistant to change?	If the intervention is uniformly effective, targeting a random sample of the population is better for spillovers than an amenable sample, and a resistant sample is better than a random sample (See Fig. 2a - 2f, Fig. 3a, and Supplementary Information). If the intervention has a declining effect, the advantages of a resistant target are lost, and a random target will tend to maximise spillovers (See Fig. 2g - 2i, Fig. 3b, and Supplementary Information). An amenable target is likely to minimize spillovers.	Mackie et al. (2015)[11] recommend an amenable target. Cloward (2016)[3] presents evidence suggesting both the tendency of development organisations to target individuals amenable to change and the limitations of such an intervention strategy. Vogt et al. (2016)[33] conducted a randomised field experiment in 122 communities in Sudan and found that participants who initially had relatively favourable attitudes towards uncut girls responded to the intervention, while participants with initially unfavourable attitudes did not.
How does homophily affect spillovers, where homophily means that people with similar preferences before the intervention tend to be linked?	For small interventions, spillovers are larger under some degree of homophily than under no homophily at all (e.g. Fig. 5d, light green line). More broadly, however, increasing homophily tends to reduce spillovers, especially for relatively large interventions (Figs. 5-6 and Supplementary Information). An intervention that targets a random sample of the population is better for inducing spillovers than either an amenable or a resistant target (Fig. 5-6). The intervention strategy should accordingly avoid selection bias and attempt to improve information flow in the population.	McPherson et al. (2001)[50] review homophily in social networks and its implications. Apicella et al. (2012)[66], Shakya et al. (2015)[16], and Migliano et al. (2017)[67] present innovative methods for identifying network structure in field settings. Centola (2011)[68] uses an online experiment to assess how homophily affects the diffusion of a low-cost innovation related to health. Cloward (2016)[3] discusses how selection bias, specifically amenable targets, can hinder programmes promoting the abandonment of cutting.
With respect to spillovers, what are the effects of a situation in which some groups use the harmful tradition to create and maintain a distinct cultural identity?	Under most conditions, using the harmful practice and its beneficial alternative as a way to define group identities and maintain group boundaries can severely constrain spillovers at the population level (Fig. 7 and Supplementary Information). The intervention strategy should strive to weaken the link between identity and tradition without provoking backlash or casting the policy maker as the salient outgroup.	Shell Duncan and Hernlund (2000)[2], Thomas (2000)[54], Gruenbaum (2001)[55], Cloward (2016)[3], Howard and Gibson (2017)[32], and Platteau et al. (2018)[12] examine the link between cutting and identity (e.g. gender, ethnicity, culture), along with the challenges this link can pose to programmes promoting abandonment.